Introduction to the Special Section on Dependable Network Computing

D.R. Avresky, *Senior Member*, *IEEE*, Jehoshua Bruck, *Fellow*, *IEEE*, and David E. Culler

DEPENDABLE Network Computing is becoming a key part of our daily economic and social life. Every day, millions of users and businesses are utilizing the Internet infrastructure for real-time electronic commerce transactions, scheduling important events, and building relationships. While network traffic and the number of users are rapidly growing, the mean-time between failures (MTTF) is surprisingly short; according to recent studies, in the majority of Internet backbone paths, the MTTF is 28 days. This leads to a strong requirement for highly dependable networks, servers, and software systems. The challenge is to build interconnected systems, based on available technology, that are inexpensive, accessible, scalable, and dependable. This special section provides insights into a number of these exciting challenges.

The following major topics are covered in the papers selected for publication:

- scalable and reliable Internet servers;
- dependable high-speed wide, local, and system area networks;
- dependable distributed mobile computing;
- reliable group communications protocols for distributed systems;
- reliable e-commerce;
- fault tolerance support in CORBA applications.

The paper "Computing in the RAIN: A Reliable Array of Independent Nodes" can be considered a position paper for this special section. Features of the RAIN system include scalability, dynamic reconfiguration, and high availability. Through software-implemented fault tolerance, the system tolerates multiple node, link, and switch faults, with no single point of failure. Three areas crucial to reliable distributed system design are examined using an integrated approach, those are communication, computing, and storage. The paper also presents applications of RAIN in designing real-time embedded systems (motivated by a joint JPL/NASA project) and Internet servers (motivated by a technology spin-off to a startup company).

- *D.R. Avresky is with the Network Computing Lab, Electrical and Computer Engineering Department, Northeastern University, 420 Dana Research Bldg., Bostson, MA 02115. E-mail: avresky@ece.neu.edu.*
- *J. Bruck is with the California Institute of Technology, Mail Code 136-93, Pasadena, CA 91125. E-mail: bruck@paradise.caltech.edu.*
- *D.E. Culler is with Computer Science Department, University of California at Berkeley, Berkeley, CA 94720. E-mail: culler@cs.berkeley.edu.*

For information on obtaining reprints of this article, please send e-mail to: tpds@computer.org, and reference IEEECS Log Number 112987.

The paper "A Protocol for Deadlock-Free Dynamic Reconfiguration in High-Speed Local Area Networks" describes an approach to dynamically reconfiguring the routing in a irregular-topology LAN comprising multiple switches and using up-down routing. The paper presents interesting work regarding dynamic reconfiguration of high speed LAN. The process affects a (usually small) region of the network. It splits the graph into correct regions, thus avoiding deadlock by modifying the virtual interregion graph. The performance of the reconfiguration techniques is compared with the static reconfiguration technique. Also, it guarantees freedom from deadlocks during the reconfiguration process without stopping user traffic during the reconfiguration, thus, it is especially suitable for systems that provide QoS guarantees. The work targets multimedia applications (audio/video streams) between pairs of nodes, where the QoS guarantees are a critical problem in the presence of failures in the network.

The paper "Implementing e-Transactions with Asynchronous Replication" presents a specification and a corresponding distributed protocol that implements electronic transactions in three tier architectures. The objective of the specification is to ensure that, when a client issues a request, there is a corresponding result computed by the application server (i.e., commit or abort), the result is committed by all database servers, and eventually delivered to the client, unless the client crashes. It extends current e-commerce technology, which does not provide high availability of the transaction outcome, but ensures at-most-once request processing, which may lead end users to obtain at-least-once transaction guarantees by retrying the transaction in the case of a failure. The specification proposed here frees the user from retrying transactions, making the transaction outcome highly available by ensuring exactly-one transaction processing.

The paper "Consensus-Based Fault-Tolerant Total Order Multicast" presents a consensus-based approach to providing total order multicast to process groups in asynchronous distributed systems. The problem is of great interest in modern distributed systems and has not been investigated as deeply as its related counterpart: Total Order Broadcast. The algorithm presented verifies two desirable properties of minimality and locality: Only the sender and the processes in the destination groups participate in the total order multicast and consensus runs only among processes in the same group. The proposed algorithm can be seen as a combination of two known protocols and it reduces to them in the two particular cases when each process is a group or when all processes are in the same group. The following

properties are used for defining the multicast primitives: uniform validity, termination, and global total order.

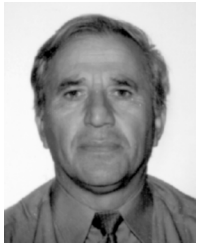
The paper "Mutable Checkpoints: A New Checkpointing Approach for Mobile Computing Systems" describes a new algorithm for distributed checkpointing in mobile computing. The algorithm is designed for distributed applications running on a combination of mobile hosts and base stations, where bandwidth, failure, and disconnection issues are paramount. The main idea is that, in a normal two-phase commit protocol for checkpointing, an application may be forced to checkpoint much more often than necessary (as we can see with a global view of all processes and a history of all communication). By adding a third form of checkpoint message, a mutable checkpoint, which essentially is a checkpoint to local disk, and causing it to be used instead of the more costly tentative checkpoint (which goes to stable storage on some server) for many of the redundant checkpoints, the authors improve the bandwidth utilization as well as overall speed of the distributed algorithm.

The paper "An Adaptive Algorithm for Tolerating Value Faults and Crash Values" presents interesting work that provides fault tolerance to CORBA applications by replicating objects and allowing the desired level of dependability to be determined at the application level. AQuA uses the Maestro/ensemble group communication system to provide reliable multicast and total ordering. Proteus, implemented on top of the Maestro/ensemble, is a flexible infrastructure for providing adaptive fault tolerance. The communication scheme, the group membership change problem, and the majority size change problem are described in great detail. At the end of the paper, the implementation is presented.

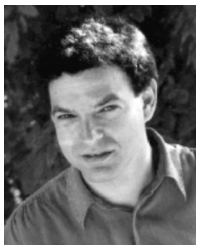
The paper "On Group Communication Support in CORBA" evaluates the use of CORBA to implement a group communication service. The motivation is to allow group communication services to be both architecture and language-independent; the authors claim that most current group communication systems assume homogeneity between the participating nodes. The use of CORBA allows different systems, potentially using different languages, to participate in a group communication service. Two group communication protocols are implemented: timewheel atomic broadcast and three-round, majority agreement group membership protocol. Evaluation is performed on three implementations: a "raw" UDP sockets-based implementation, a pure CORBA implementation, and a hybrid CORBA/UDP implementation.

In summary, we have enjoyed facilitating and editing this timely special section. We would like to thank the authors of the submitted papers for their high quality work, which made the selection process highly competitive and led to a well-balanced and interesting collection of papers. We appreciate the dedication and help of the reviewers who provided valuable feedback and insight. Finally, thanks to John Stankovic, the Editor-in-Chief, and the staff of the *IEEE Transactions on Parallel and Distributed Systems* for making this project a success.

D.R. Avresky
Jehoshua Bruck
David E. Culler



D.R. Avresky is a reader (associate professor) in the Department of Computer Science at the University of London, UK and an adjunct professor in the Department of El. and Computer Engineering at Northeastern University, Boston. Dr. Avresky's current research interests include network computing, performance analysis, cluster computing, parallel and distributed computing, fault-tolerant computing and diagnostics, embedded fault-tolerant systems, testing, and verification of protocols. He has published more than 50 papers in refereed journals and conferences. He is a co-guest editor for the *IEEE Transactions on Computers* special issue on embedded fault-tolerant systems (December 2001) and the *IEEE Micro* special issue on embedded fault-tolerant systems (September/ October 2001). He was a co-guest editor for the *IEEE Micro* special issue on embedded fault-tolerant systems (September/October 1998) and the *Journal of Supercomputing* special issue on embedded fault-tolerant systems (May 2000). He has also served as a general chair, program chair, and program committee member for numerous workshops and conferences. He edited and coauthored four books in the field of fault-tolerant parallel and distributed systems: *Dependable Network Computing* (Kluwer Academic, 1999), *Fault-Tolerant Parallel and Distributed Systems* (Kluwer Academic, 1977), *Fault-Tolerant Parallel and Distributed Systems* (IEEE CS Press, 1995), and *Hardware and Software Fault Tolerance in Parallel Computing Systems* (Simon & Schuster, 1992). He has been a consultant to several companies including Bell Labs, Tandem, and Compaq. He is a senior member of the IEEE and a member of the IEEE Computer Society.



Jehoshua (Shuki) Bruck is a professor of computation and neural systems and electrical engineering at the California Institute of Technology. His research interests include parallel and distributed computing, fault-tolerant computing, computation theory, and neural and biological systems. Dr. Bruck has extensive industrial experience, including, working with IBM for 10 years, both at the IBM Almaden Research Center and the IBM Haifa Science Center. Dr. Bruck is a cofounder and chairman of Rainfinity, a spin-off company from Caltech that is focusing on providing software for high performance reliable Internet infrastructure. He is a fellow of the IEEE.

David E. Culler received his PhD degree from the Massachusetts Institute of Technology in 1989. He is a professor of computer science at the University of California at Berkeley, where he has been a member of the faculty since 1989 and has served as vice chair for Computing and Networking. He serves on several Technical Advisory Boards and as a Faculty Computer Scientist at the National Energy Research Scientific Computing Center. He was awarded the US National Science Foundation Presidential Young Investigator in 1990 and the Presidential Faculty Fellowship in 1992. His research addresses parallel computer architecture, parallel programming languages, and high performance communication structures. He is well-known for his work on internet infrastructure, Networks of Workstations (NOW), Active Messages, Split-C, the Threaded Abstract Machine (TAM), and dataflow systems. He has published more than 70 papers in leading conferences and journals, obtained three patents, and recently completed a graduate text called, *Parallel Computer Architecture: A Hardware/Software Approach* (Morgan-Kaufmann). He has served as a co-guest editor for *IEEE Micro* (special issue on Hot Interconnects), served on several steering committees and numerous program committees for leading conferences, as well as general chair for Hot Interconnects, program chair for Hot Interconnects, and for the ACM Symposium on Parallel Algorithms and Architectures, and Technical Papers Chair for SC2001.